Privacy Preserving Adaptive Experiment Design

Jiachun Li [*1] **Kaining Shi** [*2] **David Simchi-Levi** [*1]

Abstract

Adaptive experiment is widely adopted to estimate conditional average treatment effect (CATE) in clinical trials and many other scenarios. While the primary goal in experiment is to maximize estimation accuracy, due to the imperative of social welfare, it's also crucial to provide treatment with superior outcomes to patients, which is measured by regret in contextual bandit framework. Furthermore, privacy concerns arise in clinical scenarios containing sensitive data like patients health records. Therefore, it's essential for the treatment allocation mechanism to incorporate robust privacy protection measures. In this paper, we investigate the tradeoff between loss of social welfare and statistical power of CATE estimation in contextual bandit experiment. We propose a matched upper and lower bound for the multi-objective optimization problem, and then adopt the concept of Pareto optimality to mathematically characterize the optimality condition. Furthermore, we propose *differentially private* algorithms which still matches the lower bound, showing that privacy is "almost free". Additionally, we derive the asymptotic normality of the estimator, which is essential in statistical inference and hypothesis testing.

1. Introduction

1.1. Background

The contextual bandit framework, a prominent and effective approach for sequential decision-making, is distinguished by its adaptability in progressively refining decisions based on accumulating information. This stands in contrast to reliance on static, offline datasets and batch learning methodologies. While most literature focus on developing algorithms like UCB or Thompson sampling to minimize the cumulative loss of rewards (like revenue or social welfare), it has been shown in (Lai & Robbins, 1985) that adaptive allocation strategies can surpass the efficiency of certain conventional random experimental approaches, such as Randomized Control Trials (RCTs)and has drawn much attention in recent experiment design works (Zhao 2023, Dai et al. 2023).

Considering the following motivating example of clinical trials. These trials necessitate evaluating the efficacy of new pharmaceutical interventions across diverse patient circumstances. Regret here is measured as the cumulative detriment to patient welfare, thus necessitating its minimization. This imperative becomes particularly acute in the case of rare or fatal diseases, where the goal is to allocate the most effective treatment possible to patients. The heterogeneity of patient profiles, characterized by diverse attributes such as age, gender, and genotype, significantly influences the efficacy of treatment. Therefore, it's crucial to evaluate the efficacy of drugs across varied patient profiles to identify treatments with superior therapeutic benefits while mitigating potential adverse effects for specific patient groups. This illustrates the necessity to estimate the conditional average treatment effect (CATE) (see Abrevaya et al. 2015, Fan et al. 2022, Wager & Athey 2018) in adaptive allocation assignment problems while keeping the loss of welfare, or regret at minimum. This dual focus on minimizing regret and accurately estimating CATE is central to both experimental design and contextual bandits in academic literature.

While online regret minimization and statistical inference have been extensively studied separately, the simultaneous pursuit of these objectives introduces novel complexities. This duality of purpose can result in conflicting optimal allocation strategies, as illustrated in some recent works (Simchi-Levi & Wang 2023). In specific, better accuracy of statistical inference typically necessitates broader exploration of various treatment options while a focus on minimizing regret restricts the algorithm's engagement with suboptimal arms. Moreover, the presence of patient-specific features and heterogeneous treatment effects introduce further complexity. The estimation and inference for one subgroup of patients cannot be transferred to another, and the arrival of various types of patients may be highly non-stationary, which complicates the inference process for certain subgroups due to

[*]Equal contribution [1]Laboratory for Information and Decision Systems, MIT, Cambridge, U.S. [2]Department of Statistics, The University of Chicago, Chicago, U.S.. Correspondence to: Jiachun Li <jiach334@mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

insufficient data. While the tradeoff of regret and estimation accuracy has been clearly characterized in homogeneous ATE setting, it remains unknown for the conditional average treatment case when covariates are present. We formulate it in the following question:

Question 1: *Given a budget of welfare loss (or regret), what's the best possible accuracy of estimation for CATE and how can we achieve such an accuracy?*

Privacy concerns arise in scenarios involving sensitive data types, such as healthcare records, financial information, or digital footprints. (Carlini et al. 2019, Melis et al. 2019, Niu et al. 2022). Such privacy risks also exist the estimation of Conditional Average Treatment Effect (CATE), which has the potential to inadvertently disclose sensitive individual information, including covariates, treatment statuses or outcomes. Differential Privacy (DP) has been established as a rigorous mathematical framework for defining privacy, which has gained widespread adoption in practices by major organizations such as the U.S. Census Bureau and companies like Apple and Google for data publication and analysis (Erlingsson et al. 2014, Abowd 2018). However, it is widely known that in the realms of both statistical estimation and regret minimization, "*Privacy comes at a cost*". While it is well established for DP-statistical estimation with offline, identically and independently distributed data, it's much more subtle to perform valid DP-estimation for online, potentially correlated data. Similarly, it has been a long standing problem to develop a differentially private algorithm for regret minimization in contextual bandit framework. Considering our objective to design an allocation mechanism that achieves enhanced accuracy while simultaneously minimizing regret,a DP version of our mechanism necessitates a delicate balance of these dual tasks. This leads to a critical inquiry: to what extent must one incur a "cost" to ensure privacy while striving to optimize both accuracy and regret minimization?

Question 2: *With the constraint that the experimenter need to protect the privacy of participants, is it still possible to attain the same estimation accuracy as well as social welfare loss?*

To the best of our knowledge, our work is the first one to handle these two tasks simultaneously in a DP manner.

1.2. Problem Formulation

In adaptive experiment design with heterogeneous treatment effect, there is a binary set $\mathcal{A} = \{0, 1\}$ of arms (i.e., treatments or controls) and a finite feature set $\mathcal{X} = \{X_1, X_2, \cdots, X_M\}$ with $|\mathcal{X}| = M$. Suppose n is the time horizon (or the total number of experimental units). The features at each time period follow a sequence of discrete distributions $P_X = \{P_X^t\}_{t\geq 1}$, where $P_X^t = (p_1^t, \cdots, p_M^t) \in (0,1)^M$ with $\sum_{j=1}^M p_j^t = 1, \forall t \geq 1$, representing the probability of observing experimental unit with feature X_j at time t as p_j^t. Denote $f_j(m) := \mathbb{E}\left[\sum_{1\leq t\leq m} 1_{\{x_t = X_j\}}\right] = \sum_{1\leq t\leq m} p_j^t$, which represents the expected number of occurrences of feature X_j in the first m periods, for any $1 \leq j \leq M$ and $1 \leq m \leq n$. We have the following assumption for the distribution of features.

Assumption 1.1. Seasonal Nonstationarity
(1)$\exists C_1, C_2 > 1$, s.t. $\forall 1 \leq j \leq M$, $C_1 < \frac{f_j(n)}{f_j(\frac{n}{2})} < C_2$.
(2)$f_{min}(n) := \min_{1\leq j\leq M} f_j(n) \geq \Omega(\log n)$

Intuitively, this assumption says in the first and the second half periods, every features will be expected to appear approximately same and at least $\Omega(\log n)$ times. Compared to the simplest assumption where the distribution of features for patients is stationary at each time period, our assumption greatly expands the applicability of our method. For example, different types of patients may have completely different patterns of occurrence on weekdays and weekends, or during different seasons. In this case, the simple assumption that patients' features have a stable distribution will no longer hold, but this situation may still conform to our non-stationary seasonal assumption.

Remark 1.2. Though here we denote it as non-stationarity, in fact our assumption is very mild and contains oblivious **adversarial** case. To see this, note that we allow the distribution P_t^X to be arbitrary, so for any oblivious adversarial sequence $(H_t)_{t=1}^n$ satisfying assumption 1.1, we can just set $p_i^t = 1$ for $H_t = X_i$ and $p_j^t = 0$ for any $j \neq i$ to obtain the desired sequence.

After observing the feature x_t, the experimenter will choose a treatment allocation $a_t \in \{0, 1\}$ based on policy π, and the reward of the chosen arm $r_t = r_t(a_t|x_t) \in [0, 1]$ can be observed. where $P_i(X_j)$ is the distribution of the rewards of arm i and feature X_j and P_X is the distribution of features. We define the conditional average treatment effect (CATE) of a feature x as $\Delta(X_j) := \mathbb{E}[r_t(1|X_j)] - \mathbb{E}[r_t(0|X_j)]$, for any $X_j \in \mathcal{X}$. We also denote $\sigma_{ji} = \mathbb{V}[r(i|X_j)]$ for $i = 0, 1$ and $j = 1, \cdots, M$ as the variance of reward of playing arm i when facing context X_j. In this paper, we will elaborate on $|\Delta(x)| = \Theta(1)$ for all $x \in \mathcal{X}$, which is arguably the most fundamental case in real applications. Denote all stochastic MAB instances satisfying the mentioned assumptions to constitute a feasible set $\mathcal{E}_0$.

A key index to measure the efficiency of online learning policy π is accumulative regret $\mathcal{R}(n, \pi)$, defined as the expected difference between the reward under the optimal policy and the policy π, i.e., $\mathcal{R}(n, \pi) = \mathbb{E}^\pi\left[\sum_{i=1}^n [r_i(a^*(x_i)|x_i) - r_i(a_i|x_i)]\right]$, where $a^*(x_i)$ is the optimal arm of feature x_i. In addition, an admissible adaptive estimator $\hat{\Delta}(X_j)$ maps the history $\mathcal{H}_n$ to an estimation of $\Delta(X_j)$. We use the mean square error of $\hat{\Delta}(X_j)$, i.e.,

$e\left(n, \hat{\Delta}(X_j)\right) = \mathbb{E}\left[\left(\Delta(X_j) - \hat{\Delta}(X_j)\right)^2\right]$ to measure the quality of the estimation. We define $\hat{\Delta} := \{\hat{\Delta}(X_j)\}_{1\le j\le M}$ to represent all the estimators on the gap between two arms for each feature. A design of an contextual bandit experiment can then be represented by an admissible pair $(\pi, \hat{\Delta})$. Different from traditional contextual bandit problems, the optimal design of contextual bandit experiments in this paper is solving the following minimax multi-objective optimization problem:

$$\min_{(\pi,\hat{\Delta})} \max_{\nu\in\mathcal{E}_0} \left(\mathcal{R}_\nu(n,\pi), \max_{1\le j\le M} e_\nu\left(n, \hat{\Delta}(X_j)\right)\right) \quad (1)$$

where we use the subscript ν to denote the contextual bandit instance. Eq.1 mathematically describes the two goals: minimizing the regret and the largest estimation error among all features.

The above is a rigorous mathematical description of the first question that we presented. This sets the stage for our second question, which concerns about the price of protecting privacy for both regret and CATE estimation, and how it will affect the balance between minimizing regret and estimation error. In order to rigorously address this question, we first need the following definition of differential privacy.

Definition 1.3. ((ε,δ)**-anticipating private contextual bandit algorithm**).
A bandit algorithm π is (ε,δ)-private if for two neighboring datasets $D = \{(x_i, a_i, r_i)_{i=1}^n\}$ and $D' = \{(\hat{x}_i, \hat{a}_i, \hat{r}_i)_{i=1}^n\}$ of feature, action and reward pairs that differs in at most one time step t, then for all $S \subseteq \mathcal{A}^{T-t}$:

$$\mathbb{P}\left(\pi\left(a_{t+1}, \cdots, a_n\right) \in S \mid D)\right) \le e^\varepsilon \mathbb{P}\left(\pi\left(a_{t+1}, \cdots, a_n\right) \in S \mid D')\right) + \delta,$$

where $\mathcal{A} = \{0, 1\}$ is the set of actions.

This definition is slightly different with the classical differential privacy (DP). (Shariff & Sheffet, 2018) propose a notion of "joint DP" in the context of linear contextual bandits and is later adopted by (Chen et al., 2022) as anticipating DP (ADP). The key difference of ADP is to restrict the output sets as allocations strictly after a patient of interest at time t. Such a restriction is motivated by two reasons. The first one is that following the classical DP will inevitably lead to linear regret. The second reason is that classical DP assumes that the adversary has access to the provided action a_t at time t, which is unreasonable in most adaptive experiments, as communication about (x_t, a_t, r_t) at time t is expected to be secured and the data prior to time t have no impact on the privacy of patient t because the decision making algorithm has no knowledge of x_t before time t. Therefore, only the privacy of outputs *after* time t needs to be protected. For a more detailed discussion about ADP, one can refer to (Chen et al., 2022).

Definition 1.4. ((ε,δ)**-private CATE estimator**) An admissible CATE estimator $\hat{\Delta}$ which takes a dataset $\{(x_i, a_i, r_i)\}_{i=1}^n$ as input, and output M estimations for ATE of each feature $\{\hat{\Delta}(X_i)\}_{i=1}^M$ is (ε,δ)-private if for two neighboring datasets $D = \{(x_i, a_i, r_i)_{i=1}^n\}$ and $D' = \{(\hat{x}_i, \hat{a}_i, \hat{r}_i)_{i=1}^n\}$ of feature, action and reward pairs that differs in at most one time step t, then for any measurable set $S \subset R^M$:

$$\mathbb{P}\left((\hat{\Delta}(X_1), \cdots, \hat{\Delta}(X_M) \in S \mid D\right) \le e^\varepsilon \mathbb{P}\left((\hat{\Delta}(X_1), \cdots, \hat{\Delta}(X_M) \in S \mid D'\right) + \delta.$$

Since a design of contextual bandit experiments can be represented as an admissible pair $(\pi, \hat{\Delta})$, in this paper we say a contextual bandit experiment design is (ε,δ)-private when both π and $\hat{\Delta}$ are (ε,δ)-private.

Technical Difficulties and Our Contribution

1. Elaborating on the Length of RCTs with a Regret Budget for Different Features. As claimed in (Simchi-Levi & Wang, 2023), the key idea of balancing regret and estimation error is to properly set the length of RCT. However, in our setting each feature may vary enormously in their occurrences, and it can also be highly non-stationary. Since we are interested in the worst estimation among all features, we should set the length of RCTs for all features to be the same as that of the feature with least occurrence frequency, i.e, $f_{min}(n)$. Since we don't know $f_{min}(n)$ at the beginning of experiment, by the assumption of seasonal non-stationarity, we propose an algorithm named ConSE, which divides the experiment into two phase: in the first half periods, it chooses to minimize regret while learning the frequency of occurrences $f_j(\frac{n}{2})$, and in the second half periods ConSE runs RCT for $\hat{f}_{min}(\frac{n}{2})$ periods for each feature, which is estimated from the first phase.

Another contribution of our result is the improvement of analysis compared to existing work(Simchi-Levi & Wang 2023). In particular, we develop a tighter upper bound which is tight up to constant, which helps to have a better characterization of the Pareto optimal curve for regret and estimation error. Besides, the proposed estimator in this paper is asymptotically normal, which is vital in constructing confidence interval and testing hypothesis and has been a long standing issue for adaptive experiment design literature(Simchi-Levi & Wang 2023, Dai et al. 2023, Zhao 2023). See section 3 for a more detailed discussion.

2. Privatizing Feature Information in Non-stationary Environment. Differential privacy is known to be more challenging in bandit setting due to its highly correlated data. For multi-arm bandit, algorithms based on tree mechanism proposed in (Chan et al., 2011) have been proved to be optimal up to polylog factors(see Tossou & Dimitrakakis 2016, Azize & Basu 2022, Sajed & Sheffet 2019). However, when it comes to contextual bandit, things become more complicated, as the algorithm not only needs to privatize the reward of each arm, but also the context of each patient. Most existing works focus on setting where reward

function is in a specific function class like (generalized) linear function (Hanna et al. 2022, Shariff & Sheffet 2018, Zheng et al. 2020, Chen et al. 2022). However, in clinical trials, it's risky to believe the treatment effect of one type of customer can be generalized to other types in certain way (like a linear function). Therefore, in this paper we don't assume any structure of CATE among different types of patients, which forces us to propose mechanisms different from existing literature. The second difficulty arises from the non-stationarity assumption, which has been considered in very few works. In particular, this rules out the possibility of merging different features as a whole and applying a unified mechanism.

To overcome all the difficulties mentioned, we propose a "Doubly Private" algorithm, which treats each feature separately and doubly privatize the patients' information: first of all, we use the idea of tree mechanism and divide the whole experiment into batches, where the estimation of rewards are only updated at the end of each batches. Secondly, we randomize the length of each batch to protect the context information, which is, to our best knowledge, novel in DP-contextual bandit setting. Finally, our "Doubly Private" allows experimenters to balancing regret and the estimation accuracy of CATE in any given level, and our subsequent theoretical guarantees ensure that no method can simultaneously outperform our algorithm in minimizing regret and accurately estimating CATE.

1.3. Literature Review

Adaptive Experiment Design. Experimental design is witnessing a surge in popularity across operations research, econometrics, and statistics (see, e.g., Johari et al. 2015, Bojinov et al. 2021, Bojinov et al. 2023,Xiong et al. 2023) Adaptive experimental design emerges as a particularly relevant area to our current focus(Hahn et al. 2011, Atan et al. 2019, Greenhill et al. 2020, Kato et al. 2020, Qin & Russo 2022) There are some recent works trying to demonstrate the statistical superiority of adaptive experiment compared classical non-adaptive experiment design, where the measurement of precision is the (asymptotic) variance of the estimator. In (Dai et al., 2023), they propose a measurement called Neyman regret, and show that an adaptive design with asymptotically optimal variance is equivalent to sub-linear Neyman regret, thus transforming it into a regret minimization problem. (Zhao, 2023) consider a similar setting, but adopt a competitive analysis framework.

Another emerging field is multitasking bandit problems, where minimizing regret is not the only objective (see, e.g., Yang et al. 2017, Yao et al. 2021, Zhong et al. 2021). (Erraqabi et al., 2017) also consider the tradeoff between regret and estimation error, and propose a new loss function combining these two objectives together. The most related work to this paper may be (Simchi-Levi & Wang, 2023), where they consider the tradeoff between regret and ATE estimation. We extend their framework to contextual bandit setting, derive a similar Pareto optimality characterization, and consider the additional requirement to protect patients' privacy.

Differentially Private (Contextual) Bandit Learning and Estimation. Differential privacy (Dwork et al. 2006) has emerged as gold-standard for privacy preserving data-analysis, as it ensures that the output of the data-analysis algorithm has minimum dependency on any individual datum. Differentially private variants of online learning algorithms have been successfully developed in various settings (Guha Thakurta & Smith 2013), including a private UCB-algorithm for the MAB problem (Azize & Basu 2022, Tossou & Dimitrakakis 2016) as well as UCB variations in the linear bandit settings (Hanna et al. 2022, Shariff & Sheffet 2018). These algorithms are in general motivated by techniques named "tree mechanism"(Chan et al. 2011, Dwork et al. 2010), which functions by continuously releasing aggregated statistics over a stream of T observations, introducing only $\frac{\text{polylog}(T)}{\varepsilon}$ noise in each time period, and thus leading to an added pseudo regret of order $\frac{\text{polylog}(T)}{\varepsilon}$. It was shown in Shariff & Sheffet 2018 that any ϵ-DP stochastic MAB algorithm must incur an added pseudo regret of $\Omega(\frac{K\log(T)}{\epsilon})$, and this lower bound is matched by Sajed & Sheffet 2019, using a batched elimination algorithm.

However, when it comes to DP-contextual bandit, so far there isn't a golden standard that works for general contextual bandit problems. Instead, most works focus on contextual linear bandit (Shariff & Sheffet 2018, Hanna et al. 2022,Charisopoulos et al. 2023) and adopt a relaxed definition named *joint-DP* or *anticipating-DP*. These works are in general variants of Lin-UCB (Abbasi-Yadkori et al., 2011), which is known to be optimal for contextual linear bandits. A lower bound for contextual linear bandit was proposed in Shariff & Sheffet 2018, and then was matched in Shariff & Sheffet 2018, Hanna et al. 2022 up to polylog factors. (Chen et al., 2022) consider differential privacy in dynamic pricing problem in a generalized linear model. A follow-up work (Chen et al. 2021) considers dynamic pricing in a non-parametric model and derive an upper bound of $\tilde{O}\left(T^{(d+2)/(d+4)} + \varepsilon^{-1}T^{d/(d+4)}\right)$, which is not known to be optimal.

There has been some initial work on differentially private causal inference methods. Lee & Bell 2013 proposed a privacy-preserving inverse propensity score estimator for estimating average treatment effect (ATE). Komarova & Nekipelov 2020 study the ramifications of differential privacy on the identification of statistical models and demonstrate the obstacles encountered in regression discontinuity design with privacy constraints. However, when it comes to

adaptive experiment, to our knowledge there is no similar work trying to estimating CATE privately.

2. A Warm-up: Upper and Lower Bound Without Privacy Constraint

In this section, we aim to answer the first question proposed in subsection 1.1, i.e. ***what's the best possible accuracy of estimation for CATE given a budget of regret***, by first showing a lower bound and then proposing an algorithm **ConSE** with matching upper bound. Besides, we also use this section as a warm-up to describe the technical difficulties of this problem and how to solve them, which can be helpful to understand the more complicated algorithm in section 3 with privacy constraints. In the following theorem, we provide a mini-max lower bound to explicitly show the best possible estimation accuracy with a constraint on regret budget.

Theorem 2.1. *For any admissible pair* $\left(\pi, \hat{\Delta}_n\right)$*, there always exists a hard instance* $\nu \in \mathcal{E}_0$ *such that* $e_\nu\left(n, \hat{\Delta}_n\right) \geq \Omega\left(\frac{M}{\mathcal{R}_\nu(n,\pi)}\right)$*, or in other words*

$$\inf_{(\pi,\hat{\Delta}_n)} \max_{\nu\in\mathcal{E}_0} \left[e_\nu\left(n, \hat{\Delta}_n\right) \mathcal{R}_\nu(n,\pi)\right] \geq \Omega(M).$$

Theorem 2.1 mathematically highlights the trade-off that a small regret will inevitably lead to a large error on the CATE estimation. In specific, it states that for any admissible pair $\left(\pi, \hat{\Delta}_n\right)$, there exists a hard instance $\nu \in \mathcal{E}$ such that the expected error is lower bounded by M times the inverse of the regret, i.e., $e_\nu\left(n, \hat{\Delta}_n\right) \geq \Omega\left(\frac{M}{\mathcal{R}_\nu(n,\pi)}\right)$. In particular, since $\mathcal{R}_\nu(n,\pi) = \mathcal{O}(\log(n))$ for UCB and TS algorithms, no estimators can not achieve smaller error than the order $\Omega\left(\frac{M}{\log(n)}\right)$ consistently over all the possible instances, which explicitly shows the limitation of regret optimal policies in terms of statistical power for estimating the CATE. On the other hand, if we ignore the regret and simply run random control trials (RCT), it can be easily shown that $e_\nu(n, \hat{\Delta}_n) = \max_{1\leq j\leq M} \mathbb{E}\left[\left(\hat{\Delta}_n(X_j) - \Delta(X_j)\right)^2\right] = \mathcal{O}\left(\frac{1}{f_{min}(n)}\right)$, which is the best possible accuracy one can obtain but will result in $\mathcal{O}(n)$ regret.

The above two cases can be regarded as two extreme cases (note that they don't match the lower bound), but in practice, the experimenter may want to find a balance of estimation accuracy and regret between these two extreme cases. In the following, we provide a family of algorithms named ConSE which depends on a parameter $\alpha \in [0, 1]$. A larger α leads to smaller regret budget and larger estimation error. In particular, when $\alpha = 1$, the algorithm ignores estimation error and focuses on minimizing regret. On the contrary, when $\alpha = 0$, the algorithm only focuses on minimizing estimation error. Moreover, for each given α, ConSE achieves the lower bound provided in theorem 2.1, which shows that it can attain every Pareto optimal point from one extreme case to the other (see figure 1). In the figure, the endpoints of the curve represent two extreme cases with minimum regret and estimation error. The other points on the curve characterize the tradeoff between these two objectives. Namely, this is the Pareto optimal curve for regret and estimation error. In section 4, we will have a more detailed discussion on variants of the Pareto optimal curve.

Remark 2.2. (**Intuitive example to illustrate the trade-off**) Since the two objectives presented in theorem 2.1 seems to be not conflicting at first glance, we find it necessary to provide an intuitive example here to illustrate why there is indeed a trade-off between these two objectives. Assume that there are only 2 arms with mean μ_1 and μ_2, where $\Delta = \mu_1 - \mu_2 > 0$. Now by definition, the regret is $\text{Reg} = \Delta * T(arm2)$, where $T(arm2)$ is the frequency of playing arm 2. So consider the following two tasks. The first task is to identify $\mu_1 > \mu_2$, and the second task is to estimate μ_1, μ_2 as accurate as possible. It's quite intuitive here that task 1 is strictly easier than task 2. Indeed, concentration inequalities tell us that it only takes $O(\frac{\log T}{\Delta^2})$ times for each arm to complete task one, while basic statistical lower bound tells us that to estimate μ_1, μ_2 with accuracy $\frac{1}{T^\alpha}$, it's necessary to play at least T^α times of each arm. Therefore, in order to minimize regret, one should only play each arm $O(\frac{\log T}{\Delta^2})$ times, identify that $\mu_1 > \mu_2$, and never play arm 2 again. In this case, it would lead to a very bad estimation of μ_2 with accuracy $\frac{\Delta^2}{\log T}$, but it's already necessary and sufficient for regret minimization. On the contrary, if the goal is to estimate μ_2 much more precisely with accuracy $\frac{1}{T^\alpha}$, then from the above analysis, we know that it'll inevitably lead to a regret of $O(T^\alpha)$.

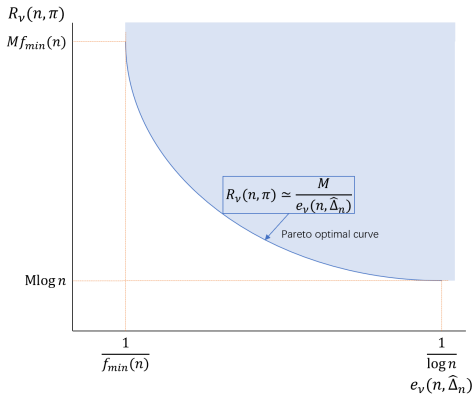

Figure 1. Pareto Optimal Curve

In the ConSE algorithm, we need the following notations.

Define three number sequences:

For $e = epoch = 1, 2, 3, ...$ and the number of total patients n, define:

$\Delta_e = 2^{-epoch}$

$R_e = \max\{\frac{32\log(16n\cdot\text{epoch}^2)}{\Delta_e^2}, \frac{8\log(8n\cdot\text{epoch}^2)}{\Delta_e}\} + 1$

$h_e = \sqrt{\frac{\log(16n\cdot\text{epoch}^2)}{2R_e}}$

Algorithm 1 ConSE

Input: α
Initialize: $S_j \leftarrow \{0,1\}$, epoch $e_j \leftarrow 0$, $r_j \leftarrow 0$, $\bar{\mu}_i^j \leftarrow 0$, $n_j \leftarrow 0$ $(i = 0, 1; j = 1, 2, ..., M)$
for $t = 1$ **to** $[\frac{n}{2}]$ **do**
  Get feature $x_t = X_{j_t} \in \mathcal{X}$
  Increment $n_{j_t} \leftarrow n_{j_t} + 1$
  if $|S_{j_t}| = 2$ **then**
    Select action $a_t \in \{0,1\}$ with equal probabilities $(\frac{1}{2}, \frac{1}{2})$ and update mean $\bar{\mu}_{a_t}^{j_t}$
    Increment $r_{j_t} \leftarrow r_{j_t} + 1$
    if $r_{j_t} \geq R_{e_{j_t}}$ **then**
      if $e_{j_t} \geq 1$ **then**
        Remove arm i from S_{j_t} if $\max\{\bar{\mu}_1^{j_t}, \bar{\mu}_2^{j_t}\} - \bar{\mu}_i^{j_t} > 2h_e$ $(i = 0, 1)$
      end if
      Increment epoch $e_{j_t} \leftarrow e_{j_t} + 1$
      Set $r_{j_t} \leftarrow 0$
      Zero means: $\bar{\mu}_i^{j_t} \leftarrow 0\ \forall i \in \{1, 2\}$
    end if
  else
    Pull the arm in S_{j_t}
  end if
  if $t = [\frac{n}{2}]$ **then**
    $\hat{f}_j = n_j (1 \leq j \leq M)$
    $T_{min} = \max\{\log n, \min\{\hat{f}_1^{1-\alpha}, \hat{f}_2^{1-\alpha}, \cdots, \hat{f}_M^{1-\alpha}\}\}$
  end if
end for
for $j = 1$ **to** M **do**
  $n_j = 0$
end for
for $t = [\frac{n}{2}] + 1$ **to** n **do**
  Get feature $x_t = X_{j_t} \in \mathcal{X}$
  Increment $n_{j_t} \leftarrow n_{j_t} + 1$
  if $n_{j_t} \leq T_{min}$ **then**
    Select action $a_t \in \{0,1\}$ with equal probabilities $(\frac{1}{2}, \frac{1}{2})$ and update mean $\bar{\mu}_{a_t}^{j_t}$
    if $n_{j_t} = T_{min}$ **then**
      Output $\hat{\Delta}(X_{j_t}) = \bar{\mu}_1^{j_t} - \bar{\mu}_0^{j_t}$
    end if
  else
    Pull the arm in S_{j_t}. (if $|S_{j_t}| = 2$, pull any arm $a_t \in S_{j_t}$)
  end if
end for

Theorem 2.3. *Let Algorithm 1 runs with any given $\alpha \in [0, 1]$. For any instance ν, the regret and estimation error are*

$$\mathcal{R}_\nu(n, \pi) \leq \mathcal{O}\left(M \max\{f_{min}(n)^{1-\alpha}, \log n\}\right),$$

$$e_\nu(n, \hat{\Delta}_n) \leq \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha}, \log n\}}\right).$$

Therefore, the product of regret and estimation error is always $\mathcal{O}(M)$, i.e.,

$$e_\nu(n, \hat{\Delta}_n)\mathcal{R}_\nu(n, \pi) \leq \mathcal{O}(M)$$

Combining the two theorems above, we can now answer ***Question 1***: Given a budget of social welfare loss $\mathcal{R}_\nu(n, \pi)$, the best possible accuracy of inference for CATE is $\mathcal{O}\left(\frac{M}{\mathcal{R}_\nu(n,\pi)}\right)$ and is attained by ConSE.

Remark 2.4. While we prove an upper bound of ConSE under non-stationary setting, the result proved in theorem 2.3 cannot be improved when the distribution of feature is stationary. This can be shown by noticing that the hard instance in lower bound in theorem 2.1 is constructed under the stationary case. That is to say, ConSE is optimal in both stationary and non-stationary settings.

Remark 2.5. Compared to previous work in bandit experiment (Simchi-Levi & Wang 2023), while the high level idea is similar, we consider an alternative estimator and improve the analysis in the proof. Specifically, in (Simchi-Levi & Wang, 2023), the upper bound is tight up to poly-log term, while in this paper the upper bound is tight up to constant. First of all, since classical bandit algorithms like UCB or TS attain regret bound of $\mathcal{O}(\log n)$, we believe that poly-log factors do matter. Besides, this improved upper bound help us have a better characterization of Pareto optimal curve that we will explain in section 3. Finally, the estimator in our algorithm is asymptotically normal, which means we can construct (asymptotic) normal confidence interval for inference and hypothesis testing, which has been a long standing issue for existing adaptive experiment design literature (Simchi-Levi & Wang 2023, Zhao 2023, Dai et al. 2023).

Table 1. Comparison with Simchi-Levi & Wang 2023.

DIFFERENCES	SIMCHI-LEVI & WANG 2023	THIS PAPER
CONTEXT	NO	YES
LOWER BOUND	$\Omega(1)$	$\Omega(M)$
UPPER BOUND	$\mathcal{O}(\log n)$	$\mathcal{O}(M)$
DIFFERENTIAL PRIVACY	NO	YES
ASYMPTOTIC NORMALITY	NO	YES

Proposition 2.6. *The estimators for all features $\hat{\Delta}(X_j)$ are unbiased, i.e., $\mathbb{E}\left[\hat{\Delta}(X_j)\right] = \Delta(X_j)$ $(\forall 1 \leq j \leq M)$. Moreover, $\hat{\Delta}(X_j)$ is asymptotically normal for any j, or formally,*

$$\lim_{n\to\infty} \sqrt{\max\{f_{min}(n)^{1-\alpha}, \log n\}}(\hat{\Delta}(X_j) - \Delta(X_j)) \rightsquigarrow \boldsymbol{N}(0, \sigma_{j0}^2 + \sigma_{j1}^2).$$

Intuitively speaking, ConSE can be divided into three steps:

Step 1. (From line 3 to 24) In the first half periods, we use Successive Elimination algorithm separately for each arm to eliminate the suboptimal arm and maintain $\log n$ regret. At the same time, we use these data to estimate the appearance frequency $f_j(\frac{n}{2})$ of each feature X_j as defined in assumption 1.

Step 2. (From line 25 to 35) At the beginning of second half periods, we run RCT $\hat{f}_{min}(\frac{n}{2})$ times for every feature, where $\hat{f}_{min}(\frac{n}{2})$ is estimated from Phase 1.

Step 3. (From line 36 to 38) Play the optimal arm for each feature in the remaining time of experiment.

Although it becomes more complicated with privacy constraints, our main goal is still to do the three steps ***privately***. A more detailed discussion will be provided in next section.

3. Privacy is Free: A Doubly-Private Algorithm for Bandit Experiment

In this section, our focus is to answer ***Question 2***, i.e., ***with the constraint that the experimenter need to protect the privacy of participants, is it still possible to attain the same estimation accuracy as well as social welfare loss?*** Roughly speaking, our answer is yes (when ε is a small, constant number, which is the most common case). In other words, we will provide a DP version of ConSE that matches the lower bound provided in theorem 2.1 for any given $\alpha \in [0, 1]$, where the meaning of α is exactly the same as in ConSE described in last section. The framework of **DP-ConSE** is quite similar to ConSE, with changes only in technical details. Due to limitation of space, we omit the precise description of DP-ConSE here and put it in appendix A.4. Instead, in the following we will provide an intuitive explanation of three steps in DP-ConSE, together with important theoretical guarantees.

Step 1. In the first half periods, we use an improved "DP Successive Elimination" algorithm in (Sajed & Sheffet, 2019) for each feature. Our goal in this phase is twofold for each feature: to identify the optimal action and estimate the frequency of occurrences (based on our non-stationary seasonal assumption) with minimal regret. For each feature we compare the **privatized** average rewards of two actions in batches. If the difference is large, we eliminate the suboptimal arm and claim that we find the optimal arm with high probability. There are two technical designs involved here. First, the length of batches increases exponentially, which strikes a balance between differential privacy protection and regret loss. Similar idea can be found in "DP Successive Elimination" algorithm (Sajed & Sheffet 2019) and widely used "tree mechanism" ((Chan et al., 2011)) in DP-bandit algorithms. Second, we use a novel technique by adding noise to the batch lengths for each feature. The reason for this is that due to the seasonal non-stationarity assumption, it's essential to run batched learning for each feature independently, and to protect the patients' feature, the length of batches should also be privatized. To the best of our knowledge, this technique has not appeared in DP-bandit literature and again highlights the difficulty of DP-contextual bandit compared to bandit setting.

After identifying the optimal action, we will continue to execute this action until the first half of the experiment is completed. After the completion of the first half, based on the occurrence frequencies of features observed, we can estimate $f_j(n)$ for each feature X_j. This helps us to decide the length of RCTs in second half periods to estimate CATE.

To make our claim valid, we first need to show that the elimination process will end in **step 1** (with high probability). This is confirmed by the following lemma.

Lemma 3.1. *Let DP-ConSE runs with any given $\alpha \in [0, 1]$ and $\varepsilon > 0$. Then w.p. $\geq 1 - \frac{1}{n}$ it holds that DP-ConSE pulls the bad arm of any feature X_j in the first half periods for at most*

$$\mathcal{O}\left((\log n_j + \log\log(1/\Delta(X_j)))\left(\frac{1}{\Delta(X_j)^2} + \frac{1}{\varepsilon\Delta(X_j)}\right)\right)$$

where n_j is the number of occurrences of the feature X_j $(1 \leq j \leq M)$.

So when ε is a small constant and n is sufficiently large, we can find the optimal arm for each feature in the first half with high probability, and the number of playing suboptimal arm is bounded. As a corollary, we can bound the regret in the first half periods as claimed.

Corollary 3.2. *For sufficiently large n, the expected pseudo regret in the first half periods of DP-ConSE is at most $\mathcal{O}\left(\left(\sum_{1\leq j\leq M}\frac{\log n}{\Delta(X_j)}\right) + \frac{M\log n}{\varepsilon}\right)$.*

Step 2. In the second half periods, our primary objective is to ensure the required accuracy of estimating the CATE. Using the estimated $f_j(n)$ from **step 1**, we can determine the length of RCTs for each feature to attain the desired accuracy. It is important to remember that we still need to add noise to the length of RCTs for the same reason as stated in **step 1**.

After **step 2**, the main task of estimating CATE is completed, and the estimation accuracy is provided in the following theorem.

Theorem 3.3. *If DP-ConSE runs with $\alpha \in [0, 1]$ and $\varepsilon > 0$, the estimate error is*

$$e(n, \hat{\Delta}) = \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}}\right).$$

Step 3. Finally, for each feature, after completing RCT phase in **step 2**, we simply play the optimal action obtained in the first half periods for the remaining patients with the aim of achieving minimum regret. The cumulative regret in the second half periods can be bounded as in the following lemma.

Lemma 3.4. *The expected regret in the second half periods of DP-ConSE is at most* $\mathcal{O}\left(\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}\sum_{1\le j\le M}\Delta(X_j)\right)$.

We have elaborated on how our algorithm strikes a balance between estimation, regret minimization and differential privacy, and to wrap things up, we have the following theorem to answer ***Question 2***. A rigorous proof can be found in appendix.

Theorem 3.5. *DP-ConSE is* $(\varepsilon, \frac{1}{n})$*-private. Moreover, let DP-ConSE runs with any given* $\alpha \in [0,1]$ *and* $\varepsilon > 0$*. The regret is*

$$\mathcal{O}\left(M\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}\right),$$

As a result, we have

$$e_\nu(n,\hat{\Delta}_n)\mathcal{R}_\nu(n,\pi) \le \mathcal{O}(M),$$

which is the same as theorem 2.3 and matches the lower bound in theorem 2.1.

Similar to the estimator without privacy constraint, we can also prove that the estimator in DP-ConSE is asymptotically normal, and more interestingly, it has the same asymptotic variance as in ConSE. This again shows that privacy is "free" in the case of statistical inference, as the Laplacian noise converges to 0 faster as $n \to \infty$ compared to Guassian variable, and thus has no impact on asymptotic variance of the estimator.

Proposition 3.6. *The estimators for all features* $\hat{\Delta}(X_j)$ *are unbiased, i.e.,* $\mathbb{E}\left[\hat{\Delta}(X_j)\right] = \Delta(X_j)$ $(\forall 1 \le j \le M)$. *Moreover,* $\hat{\Delta}(X_j)$ *is asymptotically normal for any* j*, or formally,*

$$\lim_{n\to\infty}\sqrt{\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}}(\hat{\Delta}(X_j)-\Delta(X_j)) \rightsquigarrow \boldsymbol{N}(0, \sigma_{j0}^2+\sigma_{j1}^2).$$

4. Pareto Optimal Curve

In this section, we will characterize the Pareto optimal curve of regret and estimation error, which is the standard measurement in multi-objective optimization problems and is widely adopted in multi-objective bandit literature (Simchi-Levi & Wang 2023, Zhong et al. 2021). A formal definition is provided in the following.

Definition 4.1. A pair of regret and estimation error $(x(n), y(n))$ is ***Pareto optimal*** with respect to n if there exists no algorithm which can attain a regret and estimation error pair $(\alpha(n), \beta(n))$ such that $\alpha(n) \le \mathcal{O}(x(n))$, $\beta(n) = o(y(n))$ or $\alpha(n) = o(x(n))$, $\beta(n) \le \mathcal{O}(y(n))$ as $n \to \infty$.

From the upper and lower bound we derive above in theorem 2.1, 2.3, 3.5, we know exactly what the Pareto optimal curve is.

Theorem 4.2. *The Pareto optimal curve for regret and estimation error (with or without privacy constraint) is characterized by*

$$e_\nu(n,\hat{\Delta}_n)\mathcal{R}_\nu(n,\pi) = \mathcal{O}(M),$$

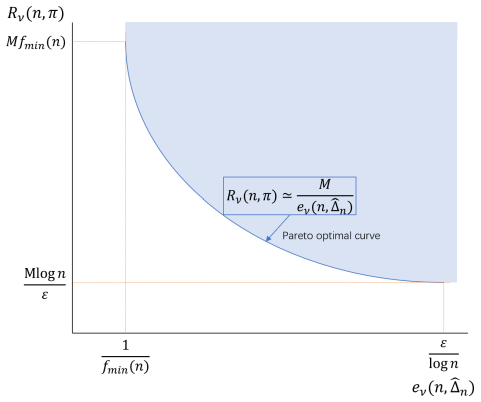

Figure 2. Pareto Optimal Curve (General)

Figure 2 shows the Pareto optimal curve in general case. We can see that the applicability of the Pareto optimal curve depends on the privacy protection indicator ϵ and the minimum feature occurrence f_{min}. As for the privacy protection parameter ϵ, we claim that it is almost "for free" in terms of the trade-off between regret and estimation error, as it does not affect the equation of our Pareto optimal curve. While here we can see that the higher the requirement for privacy protection, the shorter our optimal curve will be. This is due to the fact that the smallest possible estimation error and regret depend on ε, which is $\frac{\epsilon}{\log n}$ and $\frac{M\log n}{\varepsilon}$, respectively.

The minimum feature occurrence f_{min} is entirely determined by the (nonstationary) distribution of patient features rather than by the experimenter. Therefore, in different scenarios, we may have different optimal curves. Figure 3 shows Pareto optimal curves in different scenarios, namely when f_{min} takes different values. We can see that as f_{min} decreases gradually, the Pareto optimal curve shortens and eventually collapses into a single point. Compared to (Simchi-Levi & Wang, 2023), since they cannot deal with sub-polynomial terms, it's impossible for them to characterize the case when $f_{min} = \text{polylog}(n)$ (see figure 3). The light blue areas in different scenarios show the regions of estimation error and regret pair $\left(e_\nu(n,\hat{\Delta}_n), \mathcal{R}_\nu(n,\pi)\right)$ that algorithms may achieve, while there exists no algorithm that can attain the region below Pareto optimal curves.

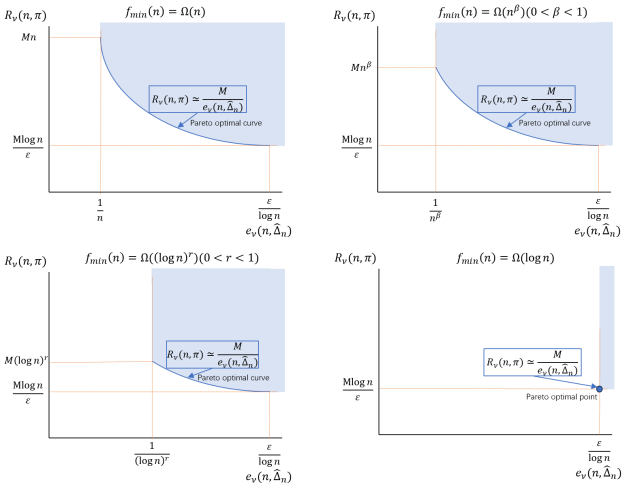

Figure 3. Pareto Optimal Curves (In Different Scenarios)

5. Experiments

In this section, we use numerical results to illustrate the theoretical findings and performance of the proposed algorithms. In particular, we focus on evaluating the algorithm ConSE in 2. First of all, we want to mention that ConSE is in fact a **meta algorithm** consisting of two phases, one for regret minimization and one for statistical estimation. For each phase, one can adopt scenario-specific algorithms to enhance potential better performance. In this paper, we adopt the batched sequential elimination for regret minimization, and RCT for CATE estimation, mainly because it's easier to privatize these two algorithms. However, when privacy is not a primary concern, one may adopt the celebrated UCB or Thopmson Sampling for regret minimization, and other estimation algorithms like adaptive neyman allocation proposed in (Zhao, 2023), or double machine learning methods with further structure assumption to improve estimation efficiency. In particular, when the outcome is assumed to have a linear structure, it was shown in (Kim et al., 2021) that doubly robust method can be adopted to capture the information in missing data. Below, we provide numerical results for estimation accuracy for both our RCT estimation and double machine learning estimation under different regret budget. We denote the mean difference estimator as MD, and double machine learning estimator as DML. The length of every experiment is 20000. For simplicity, we don't consider the heterogeneity of treatment effect among different features and assume no existence of features.

Exp 1: Normal Linear Bandit In experiment 1, we set $a_0 = [1, 0]$, $a_1 = [0, 1]$, $\theta^* = [1, 1]$, $\mu_1 = \mu_0 = 1$. The experiment results are averaged on 50 replications. From the experiment results, we can conclude that:

1. DML for predicting missing data is not helpful when experiment length n is sufficiently large.

	MD Error	DML Error
Reg=200	0.31	0.53
Reg=400	0.03	0.07
Reg=800	0.008	0.02

Table 2. Normal Linear Bandit

	MD Error	DML Error
Reg=400	0.03	4.1
Reg=800	0.008	4.0

Table 3. When regularity assumption of Linear Bandit Fails

2. The regret-estimation tradeoff always holds regardless of estimator.

Exp 2: When regularity assumption of Linear Bandit Fails

In experiment 2, we set $a_0 = [1, 0.001]$, $a_1 = [1, 0]$, $\mu_0 = 3$, $\mu_1 = 1$. In this case, the linear structure is ill-conditioned. The experiment results are averaged on 50 replications. In this experiment, MD estimator significantly outperforms DML estimator, which shows that

3. DML estimator is very sensitive in extreme cases, while MD estimator is robust.

4. MD estimator is much more efficient. The experiment of MD estimator can be finished within 1 second, while it takes 5 minutes to finish experiment of DML estimator. So to conclude, our numerical results validate the regret-estimation tradeoff and show that naive MD estimator is already efficient and powerful when experiment is long enough.

While the above experiments compare two specific estimators, it remains interesting to compare different estimation methods under various problem structures and more complex environments.

6. Concluding Remarks

In this paper, we statistically investigate the trade-off between efficiency in decision-making and estimation precision of CATE in contextual bandit experiments. We adopt the minimax multi-objective optimization framework and Pareto optimality to characterize the trade-off. We first provide a lower bound of the multi-objective optimization problem and then propose ConSE to match that lower bound. Going one step further, we consider the constraint of protecting patients' privacy and propose a differentially private version of ConSE (DP-ConSE) which still matches the lower bound, demonstrating that privacy is "almost" free. Besides, we also develop the asymptotic normality for both ConSE and DP-ConSE, which is crucial for statistical inference and hypothesis testing.

Impact Statement

This paper presents work whose goal is to advance the field of adaptive experiment design. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Abowd, J. M. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2867–2867, 2018.

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

Atan, O., Zame, W. R., and Schaar, M. Sequential patient recruitment and allocation for adaptive clinical trials. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1891–1900. PMLR, 2019.

Azize, A. and Basu, D. When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35: 32199–32210, 2022.

Bojinov, I., Rambachan, A., and Shephard, N. Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196, 2021.

Bojinov, I., Simchi-Levi, D., and Zhao, J. Design and analysis of switchback experiments. *Management Science*, 69 (7):3759–3777, 2023.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Chan, T.-H. H., Shi, E., and Song, D. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

Charisopoulos, V., Esfandiari, H., and Mirrokni, V. Robust and private stochastic linear bandits. In *International Conference on Machine Learning*, pp. 4096–4115. PMLR, 2023.

Chen, X., Miao, S., and Wang, Y. Differential privacy in personalized pricing with nonparametric demand models. *arXiv preprint arXiv:2109.04615*, 2021.

Chen, X., Simchi-Levi, D., and Wang, Y. Privacy-preserving dynamic personalized pricing with demand learning. *Management Science*, 68(7):4878–4898, 2022.

Dai, J., Gradu, P., and Harshaw, C. Clip-ogd: An experimental design for adaptive neyman allocation in sequential experiments. *arXiv preprint arXiv:2305.17187*, 2023.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 715–724, 2010.

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Erraqabi, A., Lazaric, A., Valko, M., Brunskill, E., and Liu, Y.-E. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pp. 709–717. PMLR, 2017.

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327, 2022.

Greenhill, S., Rana, S., Gupta, S., Vellanki, P., and Venkatesh, S. Bayesian optimization for adaptive experimental design: A review. *IEEE access*, 8:13937–13948, 2020.

Guha Thakurta, A. and Smith, A. (nearly) optimal algorithms for private online learning in full-information and bandit settings. *Advances in Neural Information Processing Systems*, 26, 2013.

Hahn, J., Hirano, K., and Karlan, D. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Hanna, O. A., Girgis, A. M., Fragouli, C., and Diggavi, S. Differentially private stochastic linear bandits:(almost) for free. *arXiv preprint arXiv:2207.03445*, 2022.

Johari, R., Pekelis, L., and Walsh, D. J. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Kato, M., Ishihara, T., Honda, J., and Narita, Y. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.

Kim, W., Kim, G.-S., and Paik, M. C. Doubly robust thompson sampling with linear payoffs. *Advances in Neural Information Processing Systems*, 34:15830–15840, 2021.

Komarova, T. and Nekipelov, D. Identification and formal privacy guarantees. *arXiv preprint arXiv:2006.14732*, 2020.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lee, J. Y. and Bell, D. R. Neighborhood social capital and social learning for experience attributes of products. *Marketing Science*, 32(6):960–976, 2013.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Niu, F., Nori, H., Quistorff, B., Caruana, R., Ngwe, D., and Kannan, A. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, pp. 618–633. PMLR, 2022.

Qin, C. and Russo, D. Adaptivity and confounding in multi-armed bandit experiments. *arXiv preprint arXiv:2202.09036*, 2022.

Sajed, T. and Sheffet, O. An optimal private stochastic-mab algorithm based on optimal private stopping rule. In *International Conference on Machine Learning*, pp. 5579–5588. PMLR, 2019.

Shariff, R. and Sheffet, O. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

Simchi-Levi, D. and Wang, C. Multi-armed bandit experimental design: Online decision-making and adaptive inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 3086–3097. PMLR, 2023.

Tossou, A. and Dimitrakakis, C. Algorithms for differentially private multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Xiong, R., Athey, S., Bayati, M., and Imbens, G. Optimal experimental design for staggered rollouts. *Management Science*, 2023.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Advances in Neural Information Processing Systems*, 30, 2017.

Yao, J., Brunskill, E., Pan, W., Murphy, S., and Doshi-Velez, F. Power constrained bandits. In *Machine Learning for Healthcare Conference*, pp. 209–259. PMLR, 2021.

Zhao, J. Adaptive neyman allocation. 2023.

Zheng, K., Cai, T., Huang, W., Li, Z., and Wang, L. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33: 12300–12310, 2020.

Zhong, Z., Cheung, W. C., and Tan, V. Y. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *arXiv preprint arXiv:2110.08627*, 2021.

A. Appendix

A.1. Proof to Theorem 2.1

We first consider the case that $M = 1$. In this case, we are actually talking about an ATE problem, that means no feature. And first, we start from proving the Lemma 1.1 below.

Lemma 1.1 For any given online decision-making policy π, the error of any ATE estimators can be lower bounded as follows, for any function $\phi : n \to \left[0, \frac{1}{4}\right]$ and any $u \in \mathcal{E}_0$

$$\inf_{\hat{\Delta}_n} \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu \left(\left| \hat{\Delta}_n - \Delta_\nu \right| \geq \phi(n) \right) \geq \frac{1}{2} \left[1 - \sqrt{\frac{16}{3} \phi(n)^2 \frac{\mathcal{R}_u(n, \pi)}{|\Delta_u|}} \right]$$

We define a Rademacher-like distribution as $X \sim \mathrm{Rad}(p)$ means $X = -1$ with probability p and $X = 1$ with probability $1-p$. Consider the following two bandits instance $\nu_1 = \left(\mathrm{Rad}\left(\frac{1-\xi}{2}\right), \mathrm{Rad}\left(\frac{1}{2}\right) \right)$ and $\nu_2 = \left(\mathrm{Rad}\left(\frac{1-\xi}{2}\right), \mathrm{Rad}\left(\frac{1+2\phi(t)}{2}\right) \right)$. Note that the treatment effects of ν_1 and ν_2 is $\Delta_1 = \xi$ and $\Delta_2 = \xi + 2\phi(t)$. ξ can be any number in $(0, 1]$. By such constructions and the symmetry, ν_1 and ν_2 can represent all the possible instances without loss of generality. We define the minimum distance test $\psi\left(\hat{\Delta}_t\right)$ that is associated to $\hat{\Delta}_t$ by $\psi\left(\hat{\Delta}_t\right) = \arg\min_{i=1,2} \left| \hat{\Delta}_t - \Delta_i \right|$. If $\psi\left(\hat{\Delta}_t\right) = 1$, we know that $\left| \hat{\Delta}_t - \Delta_1 \right| \leq \left| \hat{\Delta}_t - \Delta_2 \right|$. By the triangle inequality, we can have, if $\psi\left(\hat{\Delta}_t\right) = 1$,

$$\left| \hat{\Delta}_t - \Delta_2 \right| \geq |\Delta_1 - \Delta_2| - \left| \hat{\Delta}_t - \Delta_1 \right| \geq |\Delta_1 - \Delta_2| - \left| \hat{\Delta}_t - \Delta_2 \right|,$$

which yields that $\left| \hat{\Delta}_t - \Delta_2 \right| \geq \frac{1}{2} |\Delta_1 - \Delta_2| = \phi(t)$. Symmetrically, if $\psi\left(\hat{\Delta}_t\right) = 2$, we can have $| \hat{\Delta}_t - \Delta_1 | \geq \frac{1}{2} | \Delta_1 - \Delta_2 | = \phi(t)$. Therefore, we can use this to show

$$\begin{aligned}
\inf_{\hat{\Delta}_t} \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu \left(\left| \hat{\Delta}_t - \Delta_\nu \right|_2 \geq \phi(t) \right) &\geq \inf_{\hat{\Delta}_t} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} \left(\left| \hat{\Delta}_t - \Delta_i \right|_2 \geq \phi(t) \right) \\
&\geq \inf_{\hat{\Delta}_t} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} \left(\psi\left(\hat{\Delta}_t\right) \neq i \right) \\
&\geq \inf_{\psi} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} (\psi \neq i)
\end{aligned}$$

where the last infimum is taken over all tests ψ based on $\mathcal{H}_t$ that take values in $\{1, 2\}$.

$$\begin{aligned}
\inf_{\hat{\Delta}_t} \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu \left(\left| \hat{\Delta}_t - \Delta_\nu \right|_2 \geq \phi(t) \right) &\geq \inf_{\psi} \max_{i \in \{1,2\}} \mathbb{P}_{\nu_i} (\psi \neq i) \\
&\geq \frac{1}{2} \inf_{\psi} \left(\mathbb{P}_{\nu_1}(\psi = 2) + \mathbb{P}_{\nu_2}(\psi = 1) \right) \\
&= \frac{1}{2} \left[1 - \mathrm{TV}\left(\mathbb{P}_{\nu_1}, \mathbb{P}_{\nu_2} \right) \right] \\
&\geq \frac{1}{2} \left[1 - \sqrt{\frac{1}{2} \mathrm{KL}\left(\mathbb{P}_{\nu_1}, \mathbb{P}_{\nu_2} \right)} \right] \\
&\geq \frac{1}{2} \left[1 - \sqrt{\frac{8\phi(t)^2}{3\xi} \mathcal{R}_{\nu_1}(t, \pi)} \right]
\end{aligned}$$

where the equality holds due to Neyman-Pearson lemma and the third inequality holds due to Pinsker's inequality, and the fourth inequality holds due to the following:

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\nu_1}, \mathbb{P}_{\nu_2}\right) &= \sum_{s=1}^{t} \mathbb{E}_{\nu_1}\left[\mathrm{KL}\left(P_{1,A_t}, P_{2,A_t}\right)\right] \\
&= \sum_{i=1}^{2} \mathbb{E}_{\nu_1}\left[T_i(n)\right] \mathrm{KL}\left(P_{1,i}, P_{2,i}\right) \\
&= \frac{(\phi(t))^2}{\frac{1}{4} - (\phi(t))^2} \left(\mathbb{E}_{\nu_1}\left[T_i(n)\right]\right) \\
&\leq \frac{16\phi(t)^2}{3\xi} \mathcal{R}_{\nu_1}(t, \pi)
\end{aligned}$$

where we use $\mathrm{KL}\left(\mathrm{Rad}\left(\frac{1}{2}\right), \mathrm{Rad}\left(\frac{1+2\phi(t)}{2}\right)\right) = \frac{\phi(t)^2}{1/4-\phi(t)^2} \leq \frac{16\phi(t)^2}{3}$, and the last inequality holds because the history $\mathcal{H}_t$ is generated by π and $\xi\mathbb{E}_{\nu_1}\left[T_i(n)\right]$ is just the expected regret of ν_1, which is just the definition of regret. Now we finish the proof of Lemma 1.1.

Then, we can have, given policy π, and $\hat{\Delta}_n$, if $\phi(n) \leq \sqrt{\frac{3|\Delta_u|}{32\mathcal{R}_u(n,\pi)}}$ for some $u \in \mathcal{E}_0$,

$$\begin{aligned}
\max_{\nu \in \mathcal{E}_0} \mathbb{E}\left[\left|\hat{\Delta}_n - \Delta_\nu\right|^2\right] &\geq \phi(t)^2 \max_{\nu \in \mathcal{E}_0} \mathbb{P}_\nu\left(\left|\hat{\Delta}_t - \Delta_\nu\right|_2 \geq \phi(t)\right) \\
&\geq \frac{\phi(n)^2}{2}\left[1 - \sqrt{\frac{8\phi(t)^2}{3\xi}\mathcal{R}_{\nu_1}(t,\pi)}\right] \\
&\geq \frac{\phi(n)^2}{4},
\end{aligned}$$

where the second inequality holds due to Lemma 1.1.

We use $\nu_{\pi,\hat{\Delta}_n}$ to denote $\arg\max_{\nu \in \mathcal{E}_0} \mathbb{E}\left[\left|\hat{\Delta}_n - \Delta_\nu\right|^2\right]$ for any given policy π and $\hat{\Delta}_n$,

$$\begin{aligned}
\max_{\nu \in \mathcal{E}_0}\left[e_\nu\left(n, \hat{\Delta}_n\right)\mathcal{R}_\nu(n,\pi)\right] &\geq e_{\nu_{\pi,\hat{\Delta}_n}}\left(n, \hat{\Delta}_n\right)\mathcal{R}_{\nu_{\pi,\hat{\Delta}_n}}(n,\pi) \\
&\geq \frac{\phi(n)^2}{4}\mathcal{R}_{\nu_{\pi,\hat{\Delta}_n}}(n,\pi) \\
&= \Theta(1),
\end{aligned}$$

where the last equation holds because we plug in $\phi(n)$ and $\Delta_\nu = \Theta(1)$ for $\nu \in \mathcal{E}_0$. Since the above inequalities hold for any policy π and $\hat{\Delta}_n$, we finish the proof of the no feature case.

In general case, for any $1 \leq j \leq M$, we have the following:

$$\begin{aligned}
\mathcal{R}_\nu^j(n,\pi) &:= \mathbb{E}^\pi\left[\sum_{i=1}^{n} I_{\{x_i = X_j\}}\left[r_i(a^*(x_i)|x_i) - r_i\left(a_i|x_i\right)\right]\right] \\
&= \mathbb{E}\left[\mathcal{R}_\nu(n_j,\pi)\right],
\end{aligned}$$

where $n_j = \sum_{i=1}^{n} I_{\{x_i = X_j\}}$ is a random variable and $\mathbb{E}\left[\sum_{j=1}^{M}\mathcal{R}_\nu(n_j,\pi)\right] = \sum_{j=1}^{M}\mathcal{R}_\nu^j(n,\pi) = \mathcal{R}_\nu(n,\pi)$.

For using the result in no feature case, consider the following two bandits instance $\nu_1^M = \{X_j : \nu_1 | 1 \leq j \leq M\}$, $\nu_2^M = \{X_j : \nu_2 | 1 \leq j \leq M\}$, where $\nu_1 = \left(\mathrm{Rad}\left(\frac{1-\xi}{2}\right), \mathrm{Rad}\left(\frac{1}{2}\right)\right)$ and $\nu_2 = \left(\mathrm{Rad}\left(\frac{1-\xi}{2}\right), \mathrm{Rad}\left(\frac{1+2\phi(t)}{2}\right)\right)$ are introduced in the case $M = 1$.

Using the result in no feature case, we have the following:

$$\max_{\nu \in \mathcal{E}_0}\left[e_\nu\left(n, \hat{\Delta}_n(X_j)\right)\mathcal{R}_\nu(n_j,\pi)\right] \geq \Theta(1)$$

for any given n_j and $1 \leq j \leq M$.
Add their squares up, for any given $\{n_j\}_{j=1}^M$,we have

$$\max_{\nu \in \mathcal{E}_0} \left[\left(\max_{1 \leq j \leq M} e_\nu \left(n, \hat{\Delta}_n(X_j) \right) \right) \sum_{j=1}^{M} \mathcal{R}_\nu(n_j, \pi) \right] \geq \Theta(M),$$

Therefore, we have the result

$$\max_{\nu \in \mathcal{E}_0} \left[\left(\max_{1 \leq j \leq M} e_\nu \left(n, \hat{\Delta}_n(X_j) \right) \right) \mathcal{R}_\nu(n, \pi) \right] \geq \Theta(M),$$

Q.E.D.

A.2. Proof to Theorem 2.3

Firstly, we give the proof of $\mathcal{R}_\nu(n, \pi) \leq \mathcal{O}\left(M \max\{f_{min}(n)^{1-\alpha}, \log n\}\right)$ below.
Lemma 2.1 Let Algorithm 1 runs with any given $\alpha \in [0, 1]$. Then w.p. $\geq 1 - \frac{1}{n}$ it holds that Algorithm 1 pulls the bad arm of any feature X_j in the first half periods for at most

$$\mathcal{O}\left(\left(\log n_j + \log\log\left(1/\Delta(X_j)\right) \right) \frac{1}{\Delta(X_j)^2} \right)$$

where n_j is the number of occurrences of the feature X_j $(1 \leq j \leq M)$.
Proof of Lemma 2.1
Given an epoch e we denote by $\mathcal{E}_e$ the event where for all arms $a \in S$ it holds that $|\mu_a - \bar{\mu}_a| \leq h_e$ and also denote $\mathcal{E} = \bigcap_{e \geq 1} \mathcal{E}_e$. (we use $T := n_j$ represents the number of occurrences of the feature X_j and $\beta = \frac{1}{n}$ below)
First, by definition, we can calculate that:
$R_1 \geq 16 \log T$, so $R_e \geq 2R_{e-1} \geq 2^{e+3} \log T$.
Furthermore, the Hoeffding bound gives that $\Pr\left[\mathcal{E}_e\right] \geq 1 - \frac{\beta}{4e^2}$, thus $\Pr[\mathcal{E}] \geq 1 - \frac{\beta}{4}\left(\sum_{e \geq 1} e^{-2}\right) \geq 1 - \frac{1}{T}$ $(T \geq 3)$. The remainder of the proof continues under the assumption the $\mathcal{E}$ holds, and so, for any epoch e and any viable arm a in this epoch we have $|\mu_a - \bar{\mu}_a| \leq h_e$. As a result for any epoch e and any two arms a^1, a^2 we have that $|(\bar{\mu}_{a^1} - \bar{\mu}_{a^2}) - (\mu_{a^1} - \mu_{a^2})| \leq 2h_e$.

Next, we argue that under $\mathcal{E}$ the optimal arm a^* is never eliminated. Indeed, for any epoch e, we denote the arm $a_e = \operatorname{argmax}_{a \in S} \bar{\mu}_a$ and it is simple enough to see that $\bar{\mu}_{a_e} - \bar{\mu}_{a^*} \leq 0 + 2h_e$, so the algorithm doesn't eliminate a^*.

Next, we argue that, under $\mathcal{E}$, in any epoch e we eliminate all viable arms with suboptimality gap $\geq 2^{-e} = \Delta_e$. Fix an epoch e and a viable arm a with suboptimality gap $\Delta_a \geq \Delta_e$. Note that we have set parameter R_e so that

$$h_e = \sqrt{\frac{\log\left(16 \cdot e^2/\beta\right)}{2R_e}} < \sqrt{\frac{\log\left(16 \cdot e^2/\beta\right)}{2 \cdot \frac{32\log(16e^2/\beta)}{\Delta_e^2}}} = \frac{\Delta_e}{8};$$

Therefore, since arm a^* remains viable, we have that $\bar{\mu}_{\max} - \bar{\mu}_a \geq \bar{\mu}_{a^*} - \bar{\mu}_a \geq \Delta_a - (2h_e) > \Delta_e\left(1 - \frac{2}{8} - \frac{2}{8}\right) \geq \frac{\Delta_e}{2} > 2h_e$, guaranteeing that arm a is removed from S.

Lastly, fix a suboptimal arm a and let $e(a)$ be the first epoch such that $\Delta_a \geq \Delta_{e(a)}$, implying $\Delta_{e(a)} \leq \Delta_a < \Delta_{e(a)-1} = 2\Delta_e$. Using the immediate observation that for any epoch e we have $R_e \leq R_{e+1}/2$, we have that the total number of pulls of arm a is

$$\sum_{e \leq e(a)} R_e \leq \sum_{e \leq e(a)} 2^{e - e(a)} R_{e(a)} \leq R_{e(a)} \sum_{i \geq 0} 2^{-i} \leq 6\left(\frac{32\log\left(16 \cdot e(a)^2/\beta\right)}{\Delta_e^2} + \frac{8\log\left(8 \cdot e(a)^2/\beta\right)}{\Delta_e} \right)$$

The bounds $\Delta_e > \Delta_a/2, |S| \leq 2, e(a) < \log_2\left(2/\Delta_a\right)$ allow us to conclude and infer that under $\mathcal{E}$ the total number of pulls of arm a is at most

$$\log\left(2\log\left(2/\Delta_a\right)/\beta\right)\left(\frac{1024}{\Delta_a^2} + \frac{96}{\Delta_a}\right) = \mathcal{O}\left(\left(\log n_j + \log\log\left(1/\Delta(X_j)\right) \right) \frac{1}{\Delta(X_j)^2} \right)$$

We finish the proof of Lemma 2.1.
Therefore we have the following straightforward corollary.
Corollary 2.2 For sufficiently large n, the expected pseudo regret in the first half periods of Algorithm 1 is at most $\mathcal{O}\left(\sum_{1\leq j\leq M}\frac{\log n}{\Delta(X_j)}\right)$.
Actually, by using the result of lemma 2.1, for $n > \log 1/\Delta(X_j)$, we have

$$\begin{aligned}\mathcal{R}_\nu^{first}(n,\pi) &\leq \Sigma_{1\leq j\leq M}\Delta(X_j)\mathcal{O}\left(\left(\log n_j + \log\log\left(1/\Delta(X_j)\right)\right)\frac{1}{\Delta(X_j)^2}\right)\\ &\leq \mathcal{O}\left(\sum_{1\leq j\leq M}\frac{\log n}{\Delta(X_j)}\right)\end{aligned}$$

For the regret of the second half periods, noticed that with the probability $\geq 1-\frac{1}{n}$, the optimal arm would be chosen correctly in the first half periods. Therefore, the expected regret of the second half periods of DP-ConSE is:

$$\mathcal{R}_\nu^{second}(n,\pi) \leq \Sigma_{1\leq j\leq M}\Delta(X_j)\mathbb{E}\left[T_{min}\right] + \frac{1}{n}\mathcal{O}\left(n\right) = \mathcal{O}\left(\max\{f_{min}(n)^{1-\alpha},\log n\}\sum_{1\leq j\leq M}\Delta(X_j)\right)$$

Therefore, when $\Delta(X_j) = \mathcal{O}(1)$ $(\forall 1\leq j\leq M)$, we have

$$\mathcal{R}_\nu(n,\pi) = \mathcal{R}_\nu^{first}(n,\pi) + \mathcal{R}_\nu^{second}(n,\pi) \leq \mathcal{O}\left(M\max\{f_{min}(n)^{1-\alpha},\log n\}\right)$$

Secondly, we give the proof of $e_\nu(n,\hat{\Delta}_n) \leq \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha},\log n\}}\right)$ below.
Note that for any feature $X_j (1\leq j\leq M)$, we learn at least $T_j = \min\{f_j(n) - f_j(\frac{n}{2}), T_{min}\}$ periods with equal probabilities of two arms, so the MSE estimation error of feature X_j is bounded as $\mathcal{O}\left(\frac{1}{T_j}\right)$. Hence, our estimation error is bounded as $\mathcal{O}\left(\frac{1}{\min_{1\leq j\leq M}T_j}\right)$.
Now we focus on T_j.
For any $1\leq j\leq M$ and $1\leq t\leq n$, notice the characteristic function $I_{\{x_t=X_j\}}\in\{0,1\}$ and follows Bernoulli(p_j^t), we have $\mathbb{E}\left[\hat{f}_j\right] = f_j(\frac{n}{2}) = \sum_{1\leq t\leq\frac{n}{2}}p_j^t$
Therefore, by Chernoff bound (multiplicative form (relative error)), we have

$$\begin{aligned}\mathbb{P}\left(\hat{f}_j < \frac{f_j(\frac{n}{2})}{2}\right) &\leq \left(\frac{e^{-\frac{1}{2}}}{(\frac{1}{2})^{\frac{1}{2}}}\right)^{f_j(\frac{n}{2})}\\ &= \left(\sqrt{\frac{2}{e}}\right)^{f_j(\frac{n}{2})}\\ &\leq e^{-0.06f_j(\frac{n}{2})}\\ &\leq e^{-Cf_{min}(n)}\end{aligned}$$

where $C = \frac{0.06}{C_2} > 0$, the last inequality is correct due to our assumption (1).
Combining (1) and (2) in our assumption 1.1, we know $\min_{1\leq j\leq M}T_j \geq \Omega\left(\max\{f_{min}(n)^{1-\alpha},\log n\}\right)$ with at least the probability $1-Me^{-Cf_{min}(n)}$. Therefore, our estimation error is bounded as

$$e_\nu(n,\hat{\Delta}_n) \leq \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha},\log n\}}\right) + Me^{-Cf_{min}(n)}\mathcal{O}\left(1\right) = \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha},\log n\}}\right)$$

Thus, we finish the proof of theorem 2.2.

A.3. Proof to Theorem 3.3

The following proof is under the event $\bigcap_{1\le j\le M}(T_j \ge \frac{1}{2}T_{min})$, which probability is at least $1-\frac{M}{n^2}$.
Note that for any feature $X_j(1 \le j \le M)$, we learn at least $T_j = \min\{f_j(n) - f_j(\frac{n}{2}), \frac{1}{2}T_{min}\}$ periods with equal probabilities of two arms, so the MSE estimation error of feature X_j is bounded as $\mathcal{O}\left(\frac{1}{T_j}\right)$. Hence, our estimation error is bounded as $\mathcal{O}\left(\frac{1}{\min_{1\le j\le M} T_j}\right)$.
Now we focus on T_j.
For any $1 \le j \le M$ and $1 \le t \le n$, notice the characteristic function $I_{\{x_t=X_j\}} \in \{0,1\}$ and follows Bernoulli(p_j^t), we have $\mathbb{E}\left[\hat{f}_j\right] = f_j(\frac{n}{2}) = \sum_{1\le t\le \frac{n}{2}} p_j^t$
Therefore, by Chernoff bound (multiplicative form (relative error)), we have

$$\begin{aligned}\mathbb{P}\left(\hat{f}_j < \frac{f_j(\frac{n}{2})}{2}\right) &\le \left(\frac{e^{-\frac{1}{2}}}{\left(\frac{1}{2}\right)^{\frac{1}{2}}}\right)^{f_j(\frac{n}{2})} = \left(\sqrt{\frac{2}{e}}\right)^{f_j(\frac{n}{2})} \\ &\le e^{-0.06 f_j(\frac{n}{2})} \le e^{-C f_{min}(n)}\end{aligned}$$

where $C = \frac{0.06}{C_2} > 0$, the last inequality is correct due to our assumption (1).
Combining (1) and (2) in our assumption 1.1, we know $\min_{1\le j\le M} T_j \ge \Omega\left(\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}\right)$ with at least the probability $1 - Me^{-Cf_{min}(n)}$. Therefore, our estimation error is bounded as

$$\mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}}\right) + Me^{-Cf_{min}(n)}\mathcal{O}(1) + \frac{M}{n^2}\mathcal{O}(1) = \mathcal{O}\left(\frac{1}{\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}}\right)$$

.

A.4. DP-ConSE Algorithm

In this subsection, we provide the precise formulation of algorithm DP-ConSE, which is a differentially private version of ConSE provided in section 2.

Before introducing the DP-ConSE algorithm, we need to provide two notations.
Define a r.v. generator:
Given $\varepsilon > 0, \forall m > 0$, denote $Lap^+(m) = Lap_\varepsilon^+(m)$ is a random variable, satisfies:
$\forall k \ge -[m], k \in \mathcal{Z}$,
$P(Lap^+(m) = [m] + k) = \frac{e^{-\frac{\varepsilon}{2}|k|}(e^{\frac{\varepsilon}{2}}-1)}{e^{\frac{\varepsilon}{2}}+1-e^{\frac{\varepsilon}{2}[m]}}$
Define four number sequences:
For $e = epoch = 1, 2, 3, ...,$ and $\varepsilon > 0$ and the number of total patients n, define:
$\Delta_e = 2^{-epoch}$
$R_e = \max\{\frac{32\log(16n\cdot \text{epoch}^2)}{\Delta_e^2}, \frac{8\log(8n\cdot\text{epoch}^2)}{\varepsilon\Delta_e}\} + 1$
$h_e = \sqrt{\frac{\log(16n\cdot\text{epoch}^2)}{2R_e}}$
$c_e = \frac{2\log\left(8n\cdot\text{epoch}^2\right)}{R_e\varepsilon}$

Algorithm 2 DP-ConSE

Input: α, privacy-loss ε
Initialize: $S_j \leftarrow \{0,1\}$, epoch $e_j \leftarrow 0$, $r_j \leftarrow 0$, $\bar{\mu}_i^j \leftarrow 0$, $n_j \leftarrow 0$ $(i = 0, 1; j = 1, 2, ..., M)$
for $t = 1$ **to** $[\frac{n}{2}]$ **do**
  Get feature $x_t = X_{j_t} \in \mathcal{X}$
  Increment $n_{j_t} \leftarrow n_{j_t} + 1$
  if $|S_{j_t}| = 2$ **then**
    Select action $a_t \in \{0,1\}$ with equal probabilities $(\frac{1}{2}, \frac{1}{2})$ and update mean $\bar{\mu}_{a_t}^{j_t}$

Increment $r_{j_t} \leftarrow r_{j_t} + 1$
if $r_{j_t} \geq R_{e_{j_t}}^{j_t}$ **then**
if $e_{j_t} \geq 1$ **then**
Set $\widetilde{\mu}_i^{j_t} \leftarrow \bar{\mu}_i^{j_t} + \text{Lap}(\frac{2}{\varepsilon R_{e_{j_t}}})$
Remove arm i from S_{j_t} if $\max\{\widetilde{\mu}_1^{j_t}, \widetilde{\mu}_2^{j_t}\} - \widetilde{\mu}_i^{j_t} > 2h_e + 2c_e$ $(i = 0, 1)$
end if
Increment epoch $e_{j_t} \leftarrow e_{j_t} + 1$
Set $r_{j_t} \leftarrow 0$
Zero means: $\bar{\mu}_i^{j_t} \leftarrow 0 \; \forall i \in \{1, 2\}$
end if
else
Pull the arm in S_{j_t}
end if
if $t = [\frac{n}{2}]$ **then**
$\hat{f}_j = n_j (1 \leq j \leq M)$
$T_{min} = \max\{\log n, \min\{\hat{f}_1^{1-\alpha}, \hat{f}_2^{1-\alpha}, \cdots, \hat{f}_M^{1-\alpha}\}\}$
end if
end for
for $j = 1$ **to** M **do**
$T_j = Lap_{\varepsilon}^{+}(T_{min})$
$n_j = 0$
end for
for $t = [\frac{n}{2}] + 1$ **to** n **do**
Get feature $x_t = X_{j_t} \in \mathcal{X}$
Increment $n_{j_t} \leftarrow n_{j_t} + 1$
if $n_{j_t} \leq T_{j_t}$ **then**
Select action $a_t \in \{0, 1\}$ with equal probabilities $(\frac{1}{2}, \frac{1}{2})$ and update mean $\bar{\mu}_{a_t}^{j_t}$
if $n_{j_t} = T_{j_t}$ **then**
Output $\hat{\Delta}(X_{j_t}) = \bar{\mu}_1^{j_t} - \bar{\mu}_0^{j_t} + Lap(\frac{2}{\varepsilon T_{j_t}})$
end if
else
Pull the arm in S_{j_t}. (if $|S_{j_t}| = 2$, pull any arm $a_t \in S_{j_t}$)
end if
end for

A.5. Proof to Lemma 3.1, Corollary 3.2, Lemma 3.4 and Theorem 3.5

A.5.1. Proof of lemma 3.1

Given an epoch e we denote by $\mathcal{E}_e$ the event where for all arms $a \in S$ it holds that (we use $T := n_j$ represents the number of occurrences of the feature X_j and $\beta = \frac{1}{n}$ below)
(i) $|\mu_a - \bar{\mu}_a| \leq h_e$;
(ii) $|\bar{\mu}_a - \widetilde{\mu}_a| \leq c_e$;
(iii) $R_e \leq R_e^j \leq 3R_e$;
and also denote $\mathcal{E} = \bigcap_{e \geq 1} \mathcal{E}_e$.
First, by definition, we can calculate that:
$R_1 \geq \frac{16 \log T}{\epsilon}$, so $R_e \geq 2R_{e-1} \geq \frac{2^{e+3} \log T}{\epsilon}$.
Hence, $P((iii)^c) \leq 2exp\{-R_e \epsilon\} \leq 2T^{-2^{e+3}}$
Furthermore, given (iii), the Hoeffding bound, concentration of the Laplace distribution and the union bound over all arms in S_0 give that $\Pr[\mathcal{E}_e] \geq 1 - \left(\frac{\beta}{4e^2} + \frac{\beta}{4e^2} + 2T^{-2^{e+3}}\right)$, thus $\Pr[\mathcal{E}] \geq 1 - \frac{\beta}{2}\left(\sum_{e \geq 1} e^{-2}\right) - \sum_{e \geq 1} 2T^{-2^{e+3}} \geq 1 - \frac{1}{T}$ $(T \geq 3)$. The remainder of the proof continues under the assumption the $\mathcal{E}$ holds, and so, for any epoch e and any viable arm a in this epoch we have $|\widetilde{\mu}_a - \mu_a| \leq h_e + c_e$. As a result for any epoch e and any two arms a^1, a^2 we have that $|(\widetilde{\mu}_{a^1} - \widetilde{\mu}_{a^2}) - (\mu_{a^1} - \mu_{a^2})| \leq 2h_e + 2c_e$.

Next, we argue that under $\mathcal{E}$ the optimal arm a^* is never eliminated. Indeed, for any epoch e, we denote the arm $a_e = \text{argmax}_{a\in S}\widetilde{\mu}_a$ and it is simple enough to see that $\widetilde{\mu}_{a_e} - \widetilde{\mu}_{a^*} \leq 0 + 2h_e + 2c_e$, so the algorithm doesn't eliminate a^*.

Next, we argue that, under $\mathcal{E}$, in any epoch e we eliminate all viable arms with suboptimality gap $\geq 2^{-e} = \Delta_e$. Fix an epoch e and a viable arm a with suboptimality gap $\Delta_a \geq \Delta_e$. Note that we have set parameter R_e so that

$$h_e = \sqrt{\frac{\log\left(16\cdot e^2/\beta\right)}{2R_e}} < \sqrt{\frac{\log\left(16\cdot e^2/\beta\right)}{2\cdot\frac{32\log(16e^2/\beta)}{\Delta_e^2}}} = \frac{\Delta_e}{8};$$
$$c_e = \frac{\log\left(8\cdot e^2/\beta\right)}{R_e\varepsilon} < \frac{\log\left(8\cdot e^2/\beta\right)}{\varepsilon\cdot\frac{8\log(8e^2/\beta)}{\varepsilon\Delta_e}} = \frac{\Delta_e}{8}$$

Therefore, since arm a^* remains viable, we have that $\widetilde{\mu}_{\max} - \widetilde{\mu}_a \geq \widetilde{\mu}_{a^*} - \widetilde{\mu}_a \geq \Delta_a - (2h_e + 2c_e) > \Delta_e\left(1 - \frac{2}{8} - \frac{2}{8}\right) \geq \frac{\Delta_e}{2} > 2h_e + 2c_e$, guaranteeing that arm a is removed from S.

Lastly, fix a suboptimal arm a and let $e(a)$ be the first epoch such that $\Delta_a \geq \Delta_{e(a)}$, implying $\Delta_{e(a)} \leq \Delta_a < \Delta_{e(a)-1} = 2\Delta_e$. Using the immediate observation that for any epoch e we have $R_e \leq R_{e+1}/2$, we have that the total number of pulls of arm a is

$$\sum_{e\leq e(a)} R_e^j \leq 3\sum_{e\leq e(a)} R_e \leq 3\sum_{e\leq e(a)} 2^{e-e(a)}R_{e(a)} \leq 3R_{e(a)}\sum_{i\geq 0}2^{-i} \leq 6\left(\frac{32\log\left(16\cdot e(a)^2/\beta\right)}{\Delta_e^2} + \frac{8\log\left(8\cdot e(a)^2/\beta\right)}{\varepsilon\Delta_e}\right)$$

The bounds $\Delta_e > \Delta_a/2, |S| \leq 2, e(a) < \log_2\left(2/\Delta_a\right)$ allow us to conclude and infer that under $\mathcal{E}$ the total number of pulls of arm a is at most

$$3\log\left(2\log\left(2/\Delta_a\right)/\beta\right)\left(\frac{1024}{\Delta_a^2} + \frac{96}{\varepsilon\Delta_a}\right) = \mathcal{O}\left(\left(\log n_j + \log\log\left(1/\Delta(X_j)\right)\right)\left(\frac{1}{\Delta(X_j)^2} + \frac{1}{\varepsilon\Delta(X_j)}\right)\right)$$

A.5.2. Proof of corollary 3.2

By using the result of lemma 3.1, for $n > \log 1/\Delta(X_j)$, we have

$$\mathcal{R}_\nu^{first}(n,\pi) \leq \Sigma_{1\leq j\leq M}\Delta(X_j)\mathcal{O}\left(\left(\log n_j + \log\log\left(1/\Delta(X_j)\right)\right)\left(\frac{1}{\Delta(X_j)^2} + \frac{1}{\varepsilon\Delta(X_j)}\right)\right)$$
$$\leq \mathcal{O}\left(\left(\sum_{1\leq j\leq M}\frac{\log n}{\Delta(X_j)}\right) + \frac{M\log n}{\varepsilon}\right)$$

A.5.3. Proof of lemma 3.4

Noticed that with the probability $\geq 1 - \frac{1}{n}$, the optimal arm would be chosen correctly in the first half periods. Therefore, the expected regret of the second half periods of DP-ConSE is
$\mathcal{R}_\nu^{second}(n,\pi) \leq \Sigma_{1\leq j\leq M}\Delta(X_j)\mathbb{E}\left[T_j\right] + \frac{1}{n}\mathcal{O}\left(n\right) = \mathcal{O}\left(\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}\sum_{1\leq j\leq M}\Delta(X_j)\right)$

A.5.4. Proof of theorem 3.5

By adding the result of corollary 1 and lemma 2, under the condition that $\Delta(X_j) = \mathcal{O}(1)$ $(\forall 1 \leq j \leq M)$ we can easily proof that $\mathcal{R}_\nu(n,\pi) \leq \mathcal{O}\left(M\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}\right)$.
As a result, we have $e_\nu(n,\hat{\Delta}_n)\mathcal{R}_\nu(n,\pi) \leq \mathcal{O}(M)$, which is the same as theorem 2.3, and matches the lower bound in theorem 2.1.

Finally, we need to prove that the DP-ConSE is $\left(\varepsilon, \frac{1}{n}\right)$-private.
The following proof is under the event $\bigcup_{1\le j\le M}\bigcup_{e\ge 1}(R_e^j \ge R_e)$, which probability is at least $1-\frac{1}{n}$ $(n \ge 3M)$.
For any two neighboring datasets D and D', suppose D and D' are only different at time t.We discuss different cases for t as following:
(1) t is in the second half, i.e. $[\frac{n}{2}]+1 \le t \le n$;
In this case, since the probabilities of arms(actions) are not dependent of features or rewards (always $(\frac{1}{2},\frac{1}{2})$, and noticed that the output Δ and running time periods T_j are added Laplace mechanism, which are both $\frac{\epsilon}{2}$-private. Therefore, in this case, $P(D) \le e^{\epsilon}P(D')$. (2)$t$ is in the first half, i.e. $1 \le t \le [\frac{n}{2}]$;
Noticed that $|T_{min} - T'_{min}| \le 1$ and T_j and T'_j are added Laplace mechanism, which is $\frac{\epsilon}{2}$-private for the second half.
Moreover, for the first half, the mean values μs and each running periods are all added Laplace mechanism, which are $\frac{\epsilon}{2}$-private.
Therefore, by the composition theorem, $P(D) \le e^{\epsilon}P(D')$.
In conclusion, for any two neighboring datasets D and D', we have $P(D) \le e^{\epsilon}P(D') + \frac{M}{n^2} \le e^{\epsilon}P(D') + \frac{1}{n}$

A.6. Proof of Proposition 2.6 and Proposition 3.6

A.6.1. Proof of proposition 2.6

By our definition and Central Limit Theorem (CLT), we know $T_{min} = \max\{f_{min}(n)^{1-\alpha}, \log n\}$ and

$$\lim_{n\to\infty}\sqrt{T_{min}}(\hat{\Delta}(X_j) - \Delta(X_j)) \rightsquigarrow \mathbf{N}(0, \sigma_{j0}^2 + \sigma_{j1}^2)$$

.

A.6.2. Proof of proposition 3.6

We know $T_{min} = \max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}$, therefore, we have

$$\sqrt{\max\{f_{min}(n)^{1-\alpha}, \frac{\log n}{\varepsilon}\}}(\hat{\Delta}(X_j) - \Delta(X_j)) = \sqrt{T_{min}}\left((\bar{\mu}_1^j - \bar{\mu}_0^j) - (\mu_1^j - \mu_0^j) + Lap(2/\varepsilon T_j)\right)$$

To prove the above expression is asymptotically normal, we will give the proof of the three solutions below:
(1) $\frac{T_j}{T_{min}} \xrightarrow{P} 1$.
(2) $\sqrt{T_{min}}Lap(2/\varepsilon T_j) \xrightarrow{P} 0$.
(3) $\sqrt{T_{min}}\left((\bar{\mu}_1^j - \bar{\mu}_0^j) - (\mu_1^j - \mu_0^j)\right) \xrightarrow{L} \mathbf{N}(0, \sigma_{j0}^2 + \sigma_{j1}^2)$.
For the solution (1), by our definition of T_j, we know for any $\delta > 0$,

$$P\left(\left|\frac{T_j}{T_{min}} - 1\right| \ge \delta\right) \le 2\sum_{k\ge \delta T_{min}} e^{-\frac{\varepsilon}{2}|k|} \to 0$$

as $T_{min} \ge \log n \to +\infty$.
For the solution (2), by using the solution (1), we know for any $\delta \in (0,1)$,

$$P(|X| > \delta) \le P\left(\left|\frac{T_j}{T_{min}} - 1\right| \ge \delta\right) + P\left(|X| > \delta, \left|\frac{T_j}{T_{min}} - 1\right| \le \delta\right) \le P\left(\left|\frac{T_j}{T_{min}} - 1\right| \ge \delta\right) + e^{-(1-\delta)\delta\sqrt{T_{min}}\epsilon} \to 0$$

as $T_{min} \ge \log n \to +\infty$, where r.v. X follows the distribution $\sqrt{T_{min}}Lap(2/\varepsilon T_j)$.
For the solution (3), denote $S_{T_j} = T_j\left((\bar{\mu}_1^j - \bar{\mu}_0^j) - (\mu_1^j - \mu_0^j)\right)$ is the sum of T_j i.i.d differences. Similarly, we denote $S_{T_{min}}$ is the sum of T_{min} i.i.d. differences who follow the same distribution with expectation 0 and variance $\sigma_{j0}^2 + \sigma_{j1}^2$.
Obviously, by Central Limit Theorem (CLT), we have $\frac{S_{T_{min}}}{\sqrt{T_{min}}} \xrightarrow{L} \mathbf{N}(0, \sigma_{j0}^2 + \sigma_{j1}^2)$. Therefore, we only need to prove that

$$\frac{S_{T_{min}} - S_{T_j}}{\sqrt{T_{min}}} \xrightarrow{P} 0$$

By using the solution (1), we know for any $\delta, \delta' > 0$,

$$\begin{aligned} P\left(\left|\frac{S_{T_{min}} - S_{T_j}}{\sqrt{T_{min}}}\right| > \delta'\right) &\leq P\left(\left|\frac{T_j}{T_{min}} - 1\right| \geq \delta\right) + P\left(\left|\frac{S_{T_{min}} - S_{T_j}}{\sqrt{T_{min}}}\right| > \delta', \left|\frac{T_j}{T_{min}} - 1\right| \leq \delta\right) \\ &\leq P\left(\left|\frac{T_j}{T_{min}} - 1\right| \geq \delta\right) + \frac{\delta}{\delta'^2}(\sigma_{j0}^2 + \sigma_{j1}^2) \\ &\to 0 \end{aligned}$$

as $T_{min} \geq \log n \to +\infty$ and letting $\delta \to 0$, where the last inequality is because of Markov inequality.
Combining solutions (2) and (3), we can easily know the proposition 3.6 is correct.